Week 4: Model Selection and Comparative Analysis

Name: Chandan Chatragadda
Section: C
SRN: PES2UG23CS141
Course Name: Machine Learning
Submission Date: 1-09-2025

1. Introduction

This project's purpose is to build and compare classification models for the HR Attrition dataset. We do grid search for hyperparameters using two methods: A built-in library and another method by making our own gridsearch function manually.

The models we used were decision trees, KNN and logistic regression. The performance of these models was evaluated using multiple metrics.

2. Dataset Description

HR Attrition Dataset

Features: 35

Instances: 1470

Target: A binary variable indicating Attrition

3. Methodology

Key Concepts

Hyperparameter Tuning: The process of selecting the optimal settings for a model's configuration that are set prior to training, to improve its performance
Grid Search: An exhaustive method that tests all possible combinations of predefined hyperparameter values to identify the best one.

K-fold Cross-Validation: A validation method where the dataset is divided into k equal parts; the model is trained on k-1 parts and evaluated on the remaining part, repeating this process k times to assess performance consistently.

ML Pipeline

The machine learning pipeline included:

StandardScaler:- Applies normalization to the features to ensure they have a consistent scale.

SelectKBest:- Performs feature selection by choosing the top k features, where k is a hyperparameter to be optimized.

Classifiers:- Models used for prediction, including Decision Tree, k-Nearest Neighbors (k-NN), and Logistic Regression.

Implementation

Manual Implementation: Involved going through every combination of parameters using StratifiedKFold, fitting the pipeline on each training fold, and assessing performance with ROC AUC to identify the best hyperparameters.

Scikit-learn Implementation: Leveraged GridSearchCV with the same parameter grid and cross-validation approach, automating the process of training models and evaluating their performance to find the optimal parameters.

4. Results and Analysis

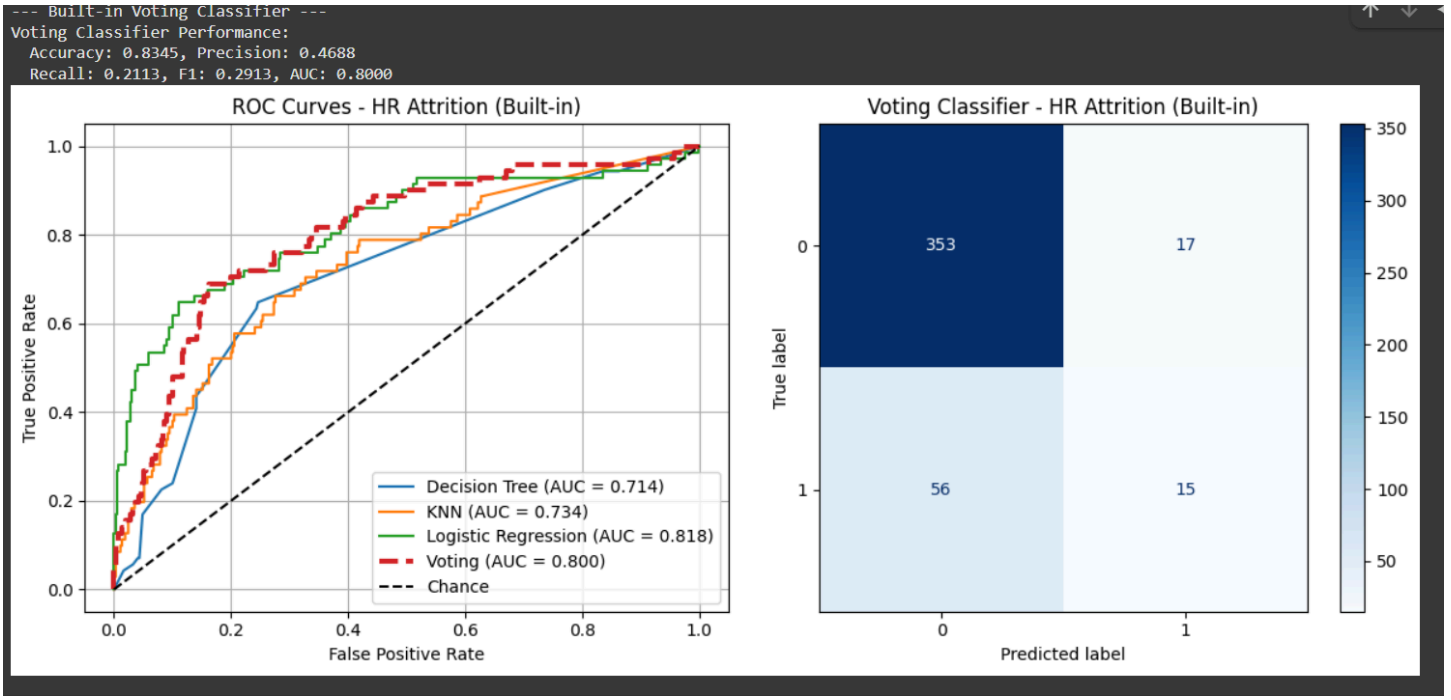
Both the manual and scikit-learn implementations produced identical results because they used the same data splits, cross-validation strategy, models, and parameter grids.

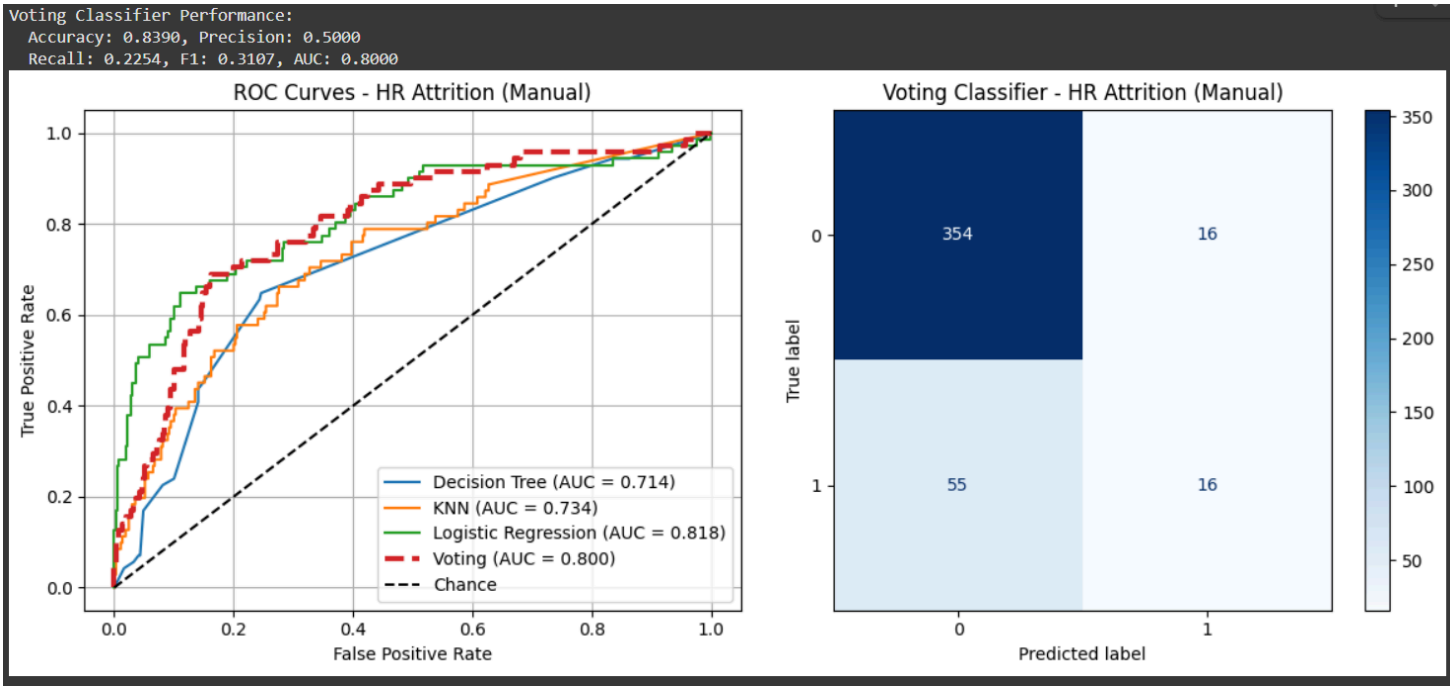
HR Attrition Dataset

Model	Accuracy	Precision	Recall	F1Score	ROC AUC
Decision Tree	0.8073	0.3478	0.2254	0.2735	0.7137
k-Nearest Neighbors	0.8277	0.4242	0.1972	0.2692	0.7340
Logistic Regression	0.8798	0.7368	0.3944	0.5138	0.8177
Voting Classifier	0.8345	0.4688	0.2113	0.2913	0.8000

The Logistic Regression and Voting Classifier models performed the best, with an ROC AUC of approximately 0.8.

5. Screenshots





6.Conculsions

Both the manual grid search and scikit-learn's GridSearchCV, when using the same parameter grid and cross-validation splits, identified the same best parameters and achieved matching performance, confirming the reliability of both methods.

In the HR Attrition dataset, Logistic Regression attained the highest ROC AUC of 0.8177, with the Voting Classifier performing nearly as well with ROC AUC of 0.8000.

More complex models or hyperparameter settings don't always guarantee better results.

Tuning the number of selected features helped improve model efficiency and accuracy by removing irrelevant or redundant data.